Arquivos Brasileiros de
Oftalmologia
PUBLICAÇÃO OFICIAL DO
CONSELHO BRASILEIRO
DE OFTALMOLOGIA

LETTERS TO THE EDITOR

Comment on: "Performance of generative large language models in answering questions from the Brazilian Retina and Vitreous Society certification exam"

Neeraj Singh[1], Monika Srivastav[2]

1. Dr. D. Y. Patil Medical College Hospital and Research Centre, Dr. D. Y. Patil Vidyapeeth (Deemed-to-be-University), Pimpri, Pune, Maharashtra 411018, India.
2. Dr. D. Y. Patil Dental College and Hospital, Dr. D. Y. Patil Vidyapeeth (Deemed-to-be-University), Pimpri, Pune, Maharashtra 411018, India.

Dear Editor:

We read with interest the study by Faneli et al., which evaluated generative large language models (LLMs), including ChatGPT-4, ChatGPT-4o, and Claude 3.5 Sonnet, using questions from the Brazilian Retina and Vitreous Society certification examination[1]. The use of a subspecialty-level question set, standardized prompting, and dual quantitative and qualitative assessments provides a clinically relevant benchmark for ophthalmic education.

The interpretation of model performance is closely linked to the decision to record only the first response as the final output. Although this design isolates baseline reasoning, it does not reflect the iterative clarification process commonly used in real-world interactions with LLMs[2]. Given that the difference in accuracy between Claude 3.5 Sonnet (72.5%) and ChatGPT-4o (66.0%) was not statistically significant, it remains uncertain whether the observed performance gap would persist under minimal refinement strategies, such as follow-up prompting or answer verification. In a certification setting, where reasoning transparency and opportunities for error correction are important, this distinction may influence how these tools are incorporated into study workflows.

The relationship between agreement metrics and accuracy also warrants careful consideration. Moderate inter-model agreement (Fleiss' κ=0.44), together with 33 questions missed by all models, suggests that errors were concentrated within specific knowledge domains rather than arising from random variability. This finding is clinically relevant because it indicates systematic limitations in higher-order clinical reasoning, particularly in areas where diagnostic and treatment accuracy remained comparatively constrained for certain models. From an educational perspective, such convergence of errors may limit the usefulness of cross-model triangulation as a validation strategy.

The qualitative assessment introduces an additional layer of complexity. Although the majority of responses were graded as "extremely correct," inter-rater agreement among graders was minimal (Fleiss' κ=0.054). This discordance suggests that qualitative ratings may have been influenced by ceiling effects or differences in interpretive thresholds rather than by a consistent assessment of explanatory depth[3]. Consequently, the apparent superiority of one model in qualitative scoring may partially reflect rating dynamics rather than substantive differences in the fidelity of clinical reasoning.

http://dx.doi.org/10.5935/0004-2749.2026-0130

Submitted for publication:
April 27, 2026

Accepted for publication:
June 9, 2026

Funding:
This study received no specific financial support.

Disclosure of potential conflicts of interest:
The authors declare no potential conflicts of interest.

Corresponding author:
Neeraj Singh
E-mail: neeraj.singh.work@proton.me

Data Availability Statement:
The datasets generated and/or analyzed during the current study are included in the manuscript.

Edited by
Editor-in-Chief: Newton Kara-Júnior
Associate Editor: Caio Vinícius S. Regatieri

The domain-specific findings further refine the interpretation of the results. Consistently higher performance in retinal pathology than that in retinal anatomy and physiology and diagnosis and treatment suggests that factual recall is more effectively captured than integrative clinical decision-making. This distinction is important because certification examinations increasingly emphasize management pathways, risk stratification, and therapeutic nuance rather than isolated knowledge recall[4].

We commend the authors for advancing the evaluation of LLMs within the context of a retina subspecialty certification framework. Future studies incorporating structured multi-turn prompting, error-clustering analyses linked to specific clinical competencies, and calibrated qualitative grading criteria may better define how these models align with the cognitive demands of specialist training and assessment.

ACKNOWLEDGMENT

The authors would like to thank Enago (www.enago.br) for the English language review.

AUTHORS' CONTRIBUTIONS:

Significant contribution to conception and design: Neeraj Singh.

Data Acquisition: Not applicable.

Data Analysis and Interpretation: Neeraj Singh.

Manuscript Drafting: Neeraj Singh; Monika Srivastav.

Significant intellectual content revision of the manuscript: Neeraj Singh; Monika Srivastav.

Final approval of the submitted manuscript: Neeraj Singh; Monika Srivastav.

Statistical analysis: Not applicable.

Obtaining funding: Not applicable.

Supervision of administrative, technical, or material support: Neeraj Singh.

Research group leadership: Neeraj Singh.

REFERENCES

1. Faneli AC, Oliveira RD, Nakayama LF, Torres RA, Muccioli C, Regatieri CV. Performance of generative large language models in answering questions from the Brazilian Retina and Vitreous Society certification exam. Arq Bras Oftalmol. 2026;89(2):e20250113.
2. Tong X, Jin B, Wang J, Xing W, Xia T, Han M. IDE: a multi-agent-driven iterative framework for dynamic evaluation of LLMs. *In:* Conference ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2025 Apr 6 [cited 2026 Jan 21];1–5. Available from: https://ieeexplore.ieee.org/document/10890123
3. Tintarev N. Measuring explanation quality – a path forward. In: Lynce N, Murano M, Vallati S, Villata F, Chesani M, Milano M, Omicini A, Dastani M (editors). ECAI 2025. 28TH European Conference of Artificial Intelligence, including 14th Conference on prestigious applications of intelligent systems, PAIS 2025. (Proceeding v. 413, p.22-29).
4. Newton WP, Handler L, Magill M. Building priorities in health & health care into ABFM's knowledge assessments. Ann Fam Med. 2022;20(3):287-9.